The Scientific Toolbox

Webis team

Overview

- Literature Research
  - Types, quality, reading, acquisition, organization
- Oral Presentations
  - Content, structure, style, talking, timing
- Scientific Writing
  - Content, structure, style, citations, plagiarism

Literature Research

Literature Research

What it is and why to do it

- Fundamental task in science
- Time-intensive but necessary
- Hardly anybody is the first on a problem

  . . . if someone is, what does that tell you?
- Don’t reinvent the wheel

- Find out if an approach to a problem is new
- Find alternative approaches or perspectives
- Widen the scope of the problem
- Obtain background information
- Obtain evidence for your or others’ claims

  . . . and similar reasons

Literature Research

Types of scientific literature (and similar)

- Textbooks, monographs
  - Theory, basics, approved techniques
- Scientific journal papers
  - Completed research lines
- Conference full papers
  - State-of-the-art research
  - Major publication type in computer science
- Conference short papers / Workshop papers
  - New ideas, ongoing research
- Technical reports
  - New ideas, ongoing research, smaller contributions
- Conference / Online tutorials
  - Easy access to basics and techniques
- Popular science magazines
  - Easy access to research lines
- Other websites
  - Anything

Literature Research

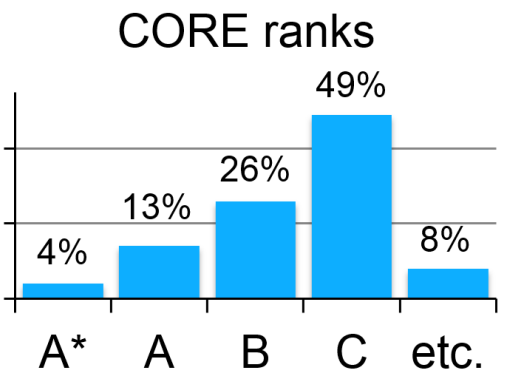
What type to prefer (in our field)

- Literature should be peer-reviewed
  - Most books, journal, conference, and workshop papers are, but not all
- Rule of thumb
  books $\succ$ journals $\succ$ conferences $\succ$ workshops $\succ$ tech reports $\succ$ magazines $\succ$ websites $\succ$ other
- . . . with exceptions like
  top conferences $\succ$ average journals

Literature Research

Assessing the “quality” of literature

- Conference and journal rankings
  - Top tier ranked A$^+$ / A* or A; B still good
  - Unranked conferences / journals may be doubtful . . . or very new
    No ranking achieves complete coverage, though.
  - One very reputable ranking is CORE
    [core.edu.au/conference-portal]



- Number of citations
  - Roughly indicates importance
  - Rather for relative comparisons within a topic
  - Remark: Newer papers naturally tend to have fewer citations
  - One resource for citation numbers is Google Scholar [scholar.google.com]
    Journals also have so-called impact factors derived from citation numbers.
- Disclaimer
  - Good and bad research appears at all places
  - Often, only reading helps . . . life is hard ;-)

Literature Research

Reading and finding literature

- Reading papers efficiently
  - Read abstract, introduction, and conclusion
  - Look at figures and tables
  - Decide whether worth reading everything
  - Read goal-driven

    Specify questions to be answered during reading.
- Finding the next paper
  - Follow promising references at the end of a paper
  - Find promising papers citing a paper
  - Learn to identify the best search terms

    Rule of thumb: As specific as possible, but as abstract as needed.
- Getting started in a seminar
  - Read the material we provide
  - Then find further literature

Literature Research

Acquiring literature

- Obtaining papers
  - Many papers simply freely available online
  - Others might be free from within a university network
  - Others might be send by authors on request
  - If neither, maybe your advisors can help
- Important sources
  - dblp for any literature related to computer science [dblp.dagstuhl.de]
  - Google Scholar or Semantic Scholar for any scientific literature
    [scholar.google.com] [semanticscholar.org]
    . . . and general web search, of course
- Accessing books
  - Check if available in the library
  - Some accessible online, for example, on Google Books [books.google.com]
    Purchasing books can make sense when of continuous importance for you.

Literature Research

Organizing literature

- Literature organization
  - Maintain notes and overview
  - “Extra” effort will pay off
- Create logical folder structure
  - Build your own view of the field
  - Logically subdivide topics, but don’t over-engineer

    For instance ./material/query-understanding/query-segmentation/ – but probably not deeper.
- Rename all PDFs consistently
  - Simplifies browsing and `grep`-ing
  - We use <author><year>-<full-title-lower-case-no-special-chars>.pdf

    As in risvik03-query-segmentation-for-web-search.pdf
- Organizing meta-information
  - Bibliographical information needed when citing literature
  - Create bibtex entries directly when organizing literature

    Very good source for computer science is dblp [`dblp.dagstuhl.de`]

Oral Presentations

Oral Presentations

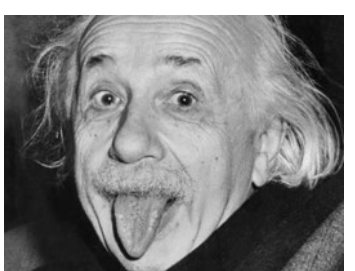
Content of a talk

- Scientific presentation is storytelling
  - Tell a coherent story with a central theme
  - Plan what points to make and how to get there
  - Make it exciting, show importance
  - Don’t be complete, be selective
    Holds for talks; different in writing
  - Avoid surprise: Clarify why you tell something

- Science needs to be understood
  - Adjust complexity to audience
  - Leave out formal things, unless really needed
    Holds for talks; probably different in writing
  - Be precise and clear
  - Introduce terms, use them consistently
  - Figures and examples help

“Sometimes reality is too complex. Stories give it a form.”

Jean-Luc Godard



“Everything should be as simple as possible, but not simpler.”

Oral Presentations

Figures

- Charts, diagrams, graphs, pictures, drawings, . . .
- Slides are visual

  Rule of thumb: No slide without a figure

- What to use figures for
  - Primary. Replace text, visually explain concepts
  - Secondary. Support your message with pictures

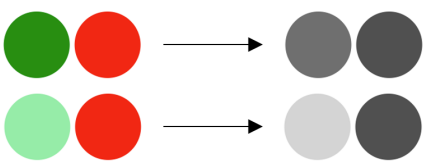
    (as often done in this presentation)
- Formatting
  - Vector graphics whenever possible
  - Others: Optimize sharpness, scale down smartly

    Don't scale > 100%; 50% is better than 53% – why?
  - Instead of squeezing or stretching the aspect ratio try to cut figures on any side
  - Think of color-blind people – contrast helps
  - Check readability of included text

"A picture is worth a 1000 words."

"Unsharpness is the mistake that even lay persons see."

Herbert Kania



Oral Presentations

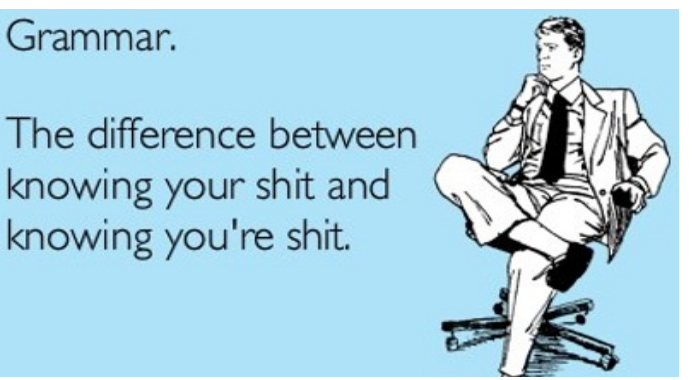
Presentation and slide structure

- Overall stucture of a presentation
  - Title slide. Title, authors, maybe date
  - Outline slide. Only for longer talks of $\gg$ 30 minutes
  - Content slides. Your story
  - Conclusion slide. Take aways, future work
    But no seperate “Thank you”-slide!
  - Maybe references. But only shown when asked for
- Structure of content slides
  - Header. Clear unique title
    Remark: Titles often not read by the audience.
  - Body. Bullet points, figures, tables, etc.
  - Footer. Title, presenter, page number, maybe “progress”
- Space for separation
  - Leave space between different slide parts
  - Leave some space to slide borders
    Sometimes clipped + it is getting harder and harder to read down here . . . isn’t it? ;-)

Oral Presentations

Style of slides

- General slide style
  - Decide what to put on a slide and what to say
  - Vary slides to maintain attention
    Larger figures here, some more text there . . .
  - Animations only when useful; use consistently
    Avoid playful ones, unless they really match your message.
  - Clarify what is from you and what from others
    Cite others' work as you do in writing (comes later).
- Text style
  - Avoid grammar and spelling errors
  - No full sentences, rather key phrases
  - AUA
    Avoid unnecessary acronyms



- Amount of text
  - Some say 7x7 maximum 7 bullet points per slide, 7 words per point
  - Others say 3x3 3 top-level points with 3 sub-points

Oral Presentations

Fonts

- Sans-serif fonts (Helvetica, Arial, . . .) much more readable on slides
- Serif fonts (Times, Garamond, . . .) maybe for example texts

- Font size – Do not mix too many on one slide
  - This text is 26pt – Maybe for titles
  - This text is 24pt
  - This text is 21pt
  - This text is 18pt – About minimum for text that should be read
  - This text is 15pt
  - This text is 12pt – Minimum for extra information that may be skipped
  - This text is 10pt
  - This text is 8pt
  - This text is 6pt – maybe for text that should not be readable ;-)

- Font shapes and colors
  - Use italics, boldface, monospace, and colors consistently
  - **And** *do not* mix `too` **many** on **a** *slide*

Oral Presentations

Talking and timing

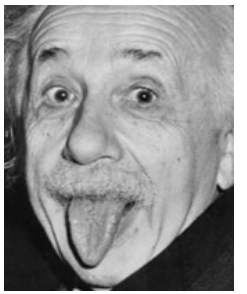
- Giving a talk
  - Match words on slides, but complement them
  - No pre-phrased sentences
  - Look at the audience, speek to everybody
  - Don’t be too formal, but be serious, avoid slang
    Jokes may be nice if you know how to use them.
- Timing
  - Use your time, but stick with given time limit
  - Expect 1.5–2 minutes per (animated) content slide
  - Rule of thumb: Audience can read slide twice
  - Leave time for questions and discussion at the end

- Practice your complete talk . . . and practice again
  - How much time do you need?
  - Do your story and slide transitions work?
  - Look for honest feedback

Scientific Writing

Scientific Writing

Content of a paper



- Most of the above hints on talks still hold
  - Science is storytelling

    Seminar: No scientific break-through expected, rather summarize and discuss.
  - Science needs to be understood
- Papers are more complete
  - Tell the whole story, avoid gaps in argumentation
  - But: Include only relevant content

    Don't expect too much prior knowledge.
  - But: No details on knowledge that can be presupposed
- Papers should be sound
  - Need to be precise more than in talks
  - Use logical arguments, from broad context to deep details
  - Formalize concepts if needed / helpful

"Don't make me think."

Steve Krug

Scientific Writing

Structure of a paper

- High-level stucture
  - Title, author information, abstract
  - Introduction
  - Usually 2–5 sections

    Related work, approach, experiments, etc.
  - Conclusion
  - References

    . . . and sometimes appendices
- Section structure
  - Often numbered subsections (2.1, 2.2, . . .)
  - If any, subsubsections unnumbered
  - Always have text introducing (sub)sections
- Section headings
  - Stick to the standard: “Introduction” is first,“Conclusion” is last, etc.
  - Short misleading headings worse than longer clear ones

The Impact of Modeling Overall Argumentation with Tree Kernels

Henning Wachsmuth
Bauhaus-Universität Weimar
Faculty of Media, Webis Group
henning.wachsmuth@uni-weimar.de

Giovanni Da San Martino
Qatar Computing Research Institute
Arabic Language Technologies
gmartino@hbku.edu.qa

Dora Kiesel
Bauhaus-Universität Weimar
Faculty of Media, VR Group
dora.kiesel@uni-weimar.de

Benno Stein
Bauhaus-Universität Weimar
Faculty of Media, Webis Group
benno.stein@uni-weimar.de

Abstract

Several approaches have been proposed to model either the explicit sequential structure of an argumentative text or its implicit hierarchical structure. So far, the adequacy of these models of overall argumentation remains unclear. This paper asks what type of structure is actually important to tackle downstream tasks in computational argumentation. We analyze patterns in the overall argumentation of texts from three corpora. Then, we adapt the idea of positional tree kernels in order to capture sequential and hierarchical argumentative structure together for the first time. In systematic experiments for three text classification tasks, we find strong evidence for the impact of both types of structure. Our results suggest that either of them is necessary while their combination may be beneficial.

1 Introduction

Argumentation theory has established a number of major argument models focusing on different aspects, such as the roles of an argument's units (Toulmin, 1958), the inference scheme of an argument (Walton et al., 2008), or the support and attack relations between arguments (Freeman, 2011). The common ground of these models is that they conceptualize an argument as a conclusion (in terms of a claim) inferred from a set of pro and con premises (reasons), which in turn may be the conclusions of other arguments. For the overall argumentation of a monological argumentative text such as the one in Figure 1(a), this results in an implicit hierarchical structure with the text's main claim at the lowest depth. In addition, the text has an explicit linguistic structure that can be seen as a regulated sequence of speech acts (van Eemeren and Grootendorst, 2004).

Figure 1: (a) Example text with five argument units, taken from the *Arg-Microtexts* corpus introduced in Section 3. (b) Graph visualization of the sequential and hierarchical overall argumentation of the text.

Figure 1(b) illustrates the interplay of the two types of overall structure in form of a tree-like graph.

Natural language processing research has largely adopted the outlined hierarchical models for mining arguments from text (Stab and Gurevych, 2014; Habernal and Gurevych, 2015; Peldszus and Stede, 2016). However, the adequacy of the resulting overall structure for downstream analysis tasks of computational argumentation has rarely been evaluated (see Section 2 for details). In fact, a computational approach that can capture patterns in hierarchical overall argumentation is missing so far. Even more, our previous work indicates that a sequential model of overall structure is preferable for analysis tasks such as stance classification or quality assessment (Wachsmuth and Stein, 2017).

In this paper, we ask and investigate what model of (monological) overall argumentation is important to tackle argumentation-related analysis tasks. To this end, we consider three corpora with fully

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2369–2379
Copenhagen, Denmark, September 7–11, 2017.

Scientific Writing

Abstract

- A concise high-level summary of the paper
- Usually 5–10 sentences

- One “approach”
  - Motivation and context (1 sentence)
  - Problem and why not solved (1–2 sentences)
  - Question addressed in the paper (1 sentence)
  - Approach with some details (2–3 sentences)
  - Evaluation, results, conclusion (1–3 sentences)

- Or in other words
  - What is the problem? Why is it a problem?
  - What is the solution? Why is it a solution to the problem?

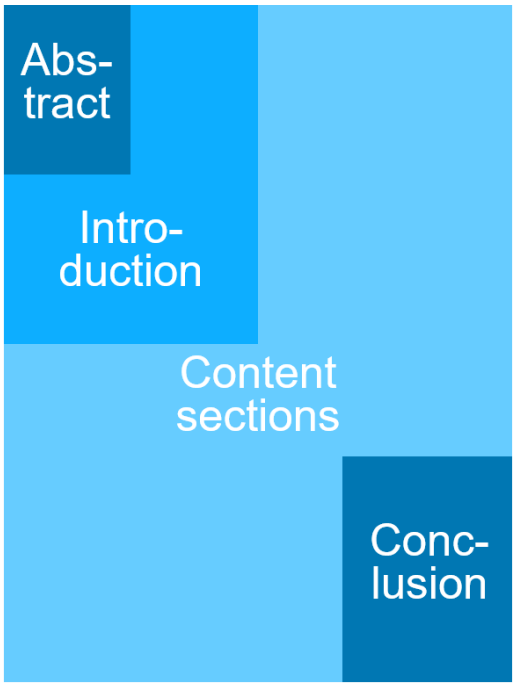
Abstract

Several approaches have been proposed to model either the explicit sequential structure of an argumentative text or its implicit hierarchical structure. So far, the adequacy of these models of overall argumentation remains unclear. This paper asks what type of structure is actually important to tackle downstream tasks in computational argumentation. We analyze patterns in the overall argumentation of texts from three corpora. Then, we adapt the idea of positional tree kernels in order to capture sequential and hierarchical argumentative structure together for the first time. In systematic experiments for three text classification tasks, we find strong evidence for the impact of both types of structure. Our results suggest that either of them is necessary while their combination may be beneficial.

Scientific Writing

Sections

- Introduction
  - The abstract in more detail
  - Tell the whole story, from context to conclusion
  - Still high-level
  - Understandable for computer scientists
- Content sections
  - The introduction in more detail
  - Elaborate on related work, concepts, models, data, approaches, experiments, and results
  - More technical, for researchers from the area
- Conclusion
  - The introduction in less detail
  - Summarize story in retrospective, give outlook
  - Semi-technical



Scientific Writing

Style

- Write clearly, unambiguously, and concise
- Don't make things complex
  (common misunderstanding)

- Some guidelines
  - Use impersonal or "we" form
  - Avoid pronouns with unclear references
  - Use explicit discourse markers, such as "because"
  - Blurring is non-scientific, such as "It could be . . . "
  - English sentences are short, one statement per sentence
  - Again: Avoid grammar and spelling errors
  - Highly recommended: *Writing for Computer Science* by Justin Zobel

Scientific Writing

Tables, figures, terms, and footnotes

- Tables and figures
  - In papers, just number increasingly
    Figure 1, 2, . . . Table 1, 2, . . . (NOT: Figure 2.1, 2.2, . . .)
  - Tables: Horizontal lines suffice
  - No included font larger than article font
  - Explain in text and in caption
- Technical terms
  - Introduce where needed, don't overformalize
  - Use well-defined terms, AIA & AUA
    Always introduce acronyms & avoid unneccessary acronyms.
  - Don't use synonyms for terms
    Reader is misled to check whether intentional differences exist.
- Footnotes
  - Only for secondary information
  - Reduced readability, should be an exception
  - Don't cite literature using footnotes

	AAE-v2	Arg-Microtexts	Web Discourse
Argument units	6089	576	1149
Avg. units/text	15.1	5.1	3.4
Min. units/text	7	3	0
Max. units/text	28	10	16
Arguments	5687	443	560
Avg. depth	2.8	2.0	0.6
Min. depth	2	1	0
Max. depth	5	4	1
Texts	402	112	340

Table 1
ments i

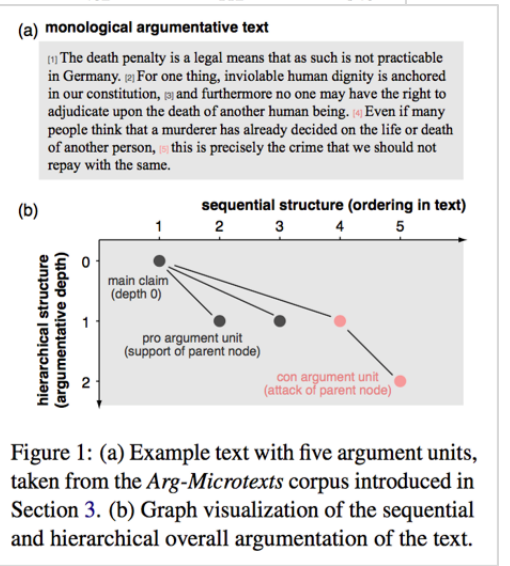


Figure 1: (a) Example text with five argument units, taken from the *Arg-Microtexts* corpus introduced in Section 3. (b) Graph visualization of the sequential and hierarchical overall argumentation of the text.

alternatives by modeling the stance of each unit towards its parent in the associated tree. This stance can be derived in all corpora.[3] All other unit and relation types from the specific models are ignored, since there is no clear mapping between them.

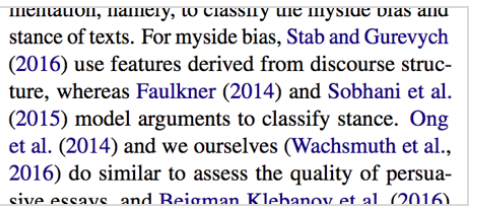
[3] Alternatively, the stance towards the main claim could be modeled. We decided against this alternative to avoid possibly wrong reinterpretations, e.g., it is unclear whether a unit that attacks its parent always supports a unit attacked by the parent.

Scientific Writing

Citations

- Citation
  - In-text reference to a bibliographic source
  - Different styles

    Acronyms [ACW17], ACL style (Ajjour et al., 2017), numbers [42], ...


mentation, namely, to classify the myside bias and stance of texts. For myside bias, Stab and Gurevych (2016) use features derived from discourse structure, whereas Faulkner (2014) and Sobhani et al. (2015) model arguments to classify stance. Ong et al. (2014) and we ourselves (Wachsmuth et al., 2016) do similar to assess the quality of persuasive essays, and Beigman Klebanov et al. (2016)

- What to cite
  - Any reuse, paraphrase, summary, or translation of content
  - Rule of thumb: Always clarify what is from you and what from others

    Also have to cite yourself if you use your own sources.
  - Better one citation too much than too few
- How to cite
  - Direct reuse. Put in quotes (shorten with [...]), give source

    Unit segmentation is "[...] the splitting of a text into argumentative segments" [ACW17].
  - Other citations. Give source close-by

    As Ajjour et al. point out, segmentation is the first task of an argument mining pipeline [ACW17].
  - Large text portions. Give source once in the beginning

    In the following paragraph, we summarize the segmentation approach of Ajjour et al. [ACW17].

Scientific Writing

References

- Bibliographical information at the end of the paper
- Exactly those references cited in the text
- Information should be complete and homogenous

- Needed meta-information
  - All literature. Author, year, title
  - Conferences/Workshops. Proceedings, pages
  - Journals. Journal name, issue, number, pages
  - Books. Edition if any, publisher
  - Only online. Give URL with access date
- Bibtex
  - LaTeX handles references automatically using bibtex

    See part on organizing literature above.

Aristotle. 2007. *On Rhetoric: A Theory of Civic Discourse* (George A. Kennedy, translator). Clarendon Aristotle series. Oxford University Press.

Beata Beigman Klebanov, Christian Stab, Jill Burstein, Yi Song, Binod Gyawali, and Iryna Gurevych. 2016. Argumentation: Content, structure, and relationship with essay quality. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 70–75. Association for Computational Linguistics.

Stefanie Brüninghaus and Kevin D. Ashley. 2003. Predicting outcomes of case based legal arguments. In *Proceedings of the 9th International Conference on Artificial Intelligence and Law*, pages 233–242.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27.

Scientific Writing

Plagiarism

- To sell another's ideas or expressions as one's own
  See en.wikipedia.org/wiki/Plagiarism
- On purpose or due to lack of giving sources
- Plagiarism is not(!) a trivial offense
  In some countries considered as crime.
- Proper citing avoids all plagiarism issues

- Consequences
  - Major cases lead to the denial of being published, graded, or worse
  - Minor cases can still negatively affect a grade or review outcomes

- Webis network [www.webis.de]
  - We do research on text reuse detection
  - See publications, shared tasks, or the tool picapica [www.picapica.org]

Take-away Messages

Scientific Writing

Conclusion

- Literature research
  - Fundamental part of scientific work
  - Literature varies in quality and suitability
  - Find, read, and organize literature efficiently

- Oral presentations and writing
  - Science is storytelling, needs to be understood
  - Best practices for content, structure, and style
  - Proper citation is a must
  - Skill comes with practice

- For you
  - Consider hints in this presentation
  - Notice that some are subjective, much is missing
  - Develop your own way